DOCUMENT RESUME

ED 157 937 TM 007 340

AUTHOR Sympson, James B.
TITLE Estimation of Latent Trait Status Using Adaptive Testing Procedures.
PUB DATE [76]
NOTE 23p.; Parts may be marginally legible due to type size

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Bayesian Statistics; *Item Analysis; *Mathematical Models; Statistical Analysis; Testing Problems
IDENTIFIERS *Adaptive Testing; Item Characteristic Curve Theory; *Latent Trait Theory

ABSTRACT

Latent trait test score theory is discussed primarily in terms of Birnbaum's three-parameter logistic model, and with some reference to the Rasch model. Equations and graphic illustrations are given for item characteristic curves and item information curves. An example is given for a hypothetical 20-item adaptive test, showing cumulative results of ability estimates for one candidate under four methods: maximum likelihood estimates; weighted by likelihoods estimates; simplified Bayesian estimates; and Owen's normal ogive Bayesian estimates. Finally, information curves are given comparing a hypothetical 20-item adaptive test with information curves for a test with a broad rectangular item-difficulty distribution and a test with a peaked item-difficulty distribution. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED157937

TM007 340

Estimation of Latent Trait Status Using Adaptive Testing Procedures

by
James B. Sympson
University of Minnesota

During the last few years, latent trait theory has become increasingly important as a theoretical foundation for the practice of psychological and educational assessment. This has been partly due to shortcomings inherent in classical test theory (see the chapter by Lumsden in Annual Review of Psychology, 1976) and partly due to recent developments in testing practice. In particular, when "adaptive" or "individualized" testing is desired, latent trait theory provides a uniquely useful conceptual scheme for guiding test design and test scoring procedures.

Latent trait theories are characterized by a mathematical model that relates the probability of occurrence of a particular response class (e.g., a "correct" response) in the presence of a particular stimulus (e.g., a test item) to a person's position on one or more metric dimensions. The graph of the function that relates probability of a particular response class to a person's status on these dimension(s) can be referred to as a response-characteristic curve.

Both univariate and multivariate latent trait models have been proposed. The univariate models (e.g., Lord, 1952; Rasch, 1960; Birnbaum, 1968; Bock, 1972) assume that response probabilities are related to the relative positions of persons and stimuli on a single metric dimension. Multivariate models (e.g., Samejima, 1974; Christoffersson, 1975) allow for the possibility of several latent dimensions.

Latent Trait Theory and the Objectives of Measurement

Before proceeding, I would like to address certain issues that trouble many people when they first encounter latent trait theory. It is reasonable to ask, "Why should I bother with an approach to testing that involves inferred latent traits, when what I'm really interested in is (choose one: predicting some criterion accurately; content validity; criterion-referenced measurement)?" While I do not have time to present extensive supporting arguments here, the following comments provide the essence of my response to these questions.

First, the adoption of latent trait theory as a guide to test construction and scoring practices does not require a belief in the "existence" of unobservable traits that control human behavior. Empirically, it is sufficient to inquire whether peoples' responses

This research was supported by Contract N0014-76-C-0243, NR 150-382, with the Personnel and Training Research Programs, Office of Naval Research, David J. Weiss, Principal Investigator.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
James B. Sympson
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

to test stimuli can be well predicted on the basis of such a model. The postulated dimensions of latent trait theory can be viewed as quantitative variables that we create by calibrating and scoring test items in a certain way. Needless to say, it is our hope that the variables we create will provide a convenient basis for designing testing procedures and will lead to increased predictive accuracy in scientific and practical applications.

In many situations tests are developed and applied with the sole intention of predicting performance in a criterion setting. The introduction of intervening variables (latent traits) might seem unnecessary when one is only interested in obtaining a high degree of relationship between test scores and criterion scores. I have two comments to make regarding this issue. First, estimates of latent trait status can themselves be viewed as a particular variety of test score. Such scores may or may not have higher predictive validity than more conventional test scores. This is an empirical question. However, even if predictive validity is not increased through the use of latent trait scores, it may still be advantageous to adopt a latent trait approach if the testing process can be made more efficient as a result (e.g., through adaptive testing procedures).

Second, it is my contention that test development for the purpose of criterion prediction is always based upon an implicit structural model. No one chooses items at random from all conceivable item domains. We try out items with certain kinds of content and never consider using other kinds of content. We also pay careful attention to generating items that have difficulty levels or endorsement rates (i.e., p-values) that are not too extreme in the population to be tested. This is done so that item-criterion correlations will not be unduly restricted. Such procedures suggest the existence of an implicit structural model.

Trying certain types of items, and not others, implies that certain types of inter-person differences exist and are related to criterion performance, while others are not. More generally, any conceptual scheme for classifying test items implies a corresponding set of response variables that can be generated when the items are administered. In selecting items for criterion prediction we indicate the type of response variable(s) we think to be related to the criterion.

A concern about item difficulties and endorsement rates implies that the probability of a given response to an item is a function of status on the relevant response variable(s). If such probabilities were not a function of status on the response variables, an item would have the same p-value in each and every population we could conceive of and there would be no need to match item difficulties to the population we plan to test.

A latent trait approach to test construction and scoring provides a formal vehicle for elaborating structural models and encourages the test developer to bring implicit structural assumptions out into the

open. Explicit structural models can serve to guide our test construction efforts and aid in the interpretation of empirical results.

With regard to content validity and criterion-referenced measurement, it is important to keep in mind that the testing situation never constitutes the entire domain of interest. Our implicit objective in pursuing content validity and in implementing criterion-referenced measurement is to make more accurate inferences about a person's potential for performance in a hypothetical task domain (Cronbach, 1971, p. 452; Glaser & Nitko, 1971, p. 653). This hypothetical task domain, though it is not observable in its entirety, is always carefully defined in terms of performance objectives or item content. Test items are generated that represent the domain, and responses to these items are used as a basis for making inferences about domain performance.

Some individuals protest such a view and argue that in criterion-referenced measurement the test stimuli *are* the criterion tasks of interest and that no further task domain is intended or implied. If this were in fact the case, we could develop a criterion-referenced test for, say, aircraft mechanics and have students remain in technical school for the entire duration of their enlistment. If the tasks in the test *are* the criterion tasks of interest, then performance on the test would constitute a sufficient basis for writing ERs (efficiency reports) and determining who should be promoted. There would be no need to have people actually go into the field and work on operational aircraft. The fact that criterion-referenced tests are being developed for use in field settings does not defuse this obviously absurd conclusion. Unless we include each and every relevant task completed by an aircraft mechanic during each and every work day in the "test", we are necessarily making inferences about a larger task domain from a sample of person-stimulus interactions drawn from the domain.

What is the nature of the hypothetical task domain in achievement testing? Given the emphasis on generating items that fulfill various content or process requirements and the emphasis on learning hierarchies (i.e., difficulty hierarchies) that permeates educational testing, I would suggest that such task domains can be described in terms of a multidimensional structural model. Whenever test stimuli can be clustered with regard to common content or process and then ordered within a cluster, there is a definite possibility that a latent trait approach to achievement testing will be useful.

In recent years the distinction between norm-referenced and criterion-referenced measurement has been widely discussed. An important fact to keep in mind is that this distinction properly applies to the type of information available from test scores, not to test content or the testing procedure itself (Hambleton & Novick, 1973, p. 162). This is an important fact because, as it turns out, estimates of latent trait status can provide information about both inter-person differences (norm-referenced interpretations) and intra-

person response probabilities for tasks drawn from a task domain (criterion-referenced interpretations).

An estimate of an individual's latent trait status can be converted to a centile rank or standard score relative to any norm group previously tested using the latent trait procedure. This same latent trait estimate, when considered in conjunction with the latent trait parameters of a test item (i.e., a task sample) that has been previously calibrated, allows us to generate the probability of occurrence of a given response class (e.g., a "correct" response) in the presence of the item. The fact that latent trait theory can provide both norm-referenced and criterion-referenced interpretations of test performance indicates that the current schism between psychological and educational testing may be narrowed considerably in the years to come.

Estimating Latent Trait Status

Hopefully, at this point I have convinced you that latent trait theory has a wide range of potential applications. If so, then procedures for estimating latent trait status should also be of interest to you. In the following discussion I will present four methods for obtaining estimates of latent trait status and will argue that the accuracy of such estimates can be improved through the use of adaptive testing procedures.

In my example, I will deal with a latent trait model in which only two response classes are considered, a *keyed* response and a *non-keyed* response, and where the probability of occurrence of each response class is a function of a single latent dimension. This model might be applicable to a test that has been constructed to maximize internal consistency (Nunnally, 1967, pp. 254-268) and in which items are scored dichotomously. The model would not be suitable for tests that involve a multidimensional item structure, but the principles of latent trait estimation that I will present are generalizable to such cases.

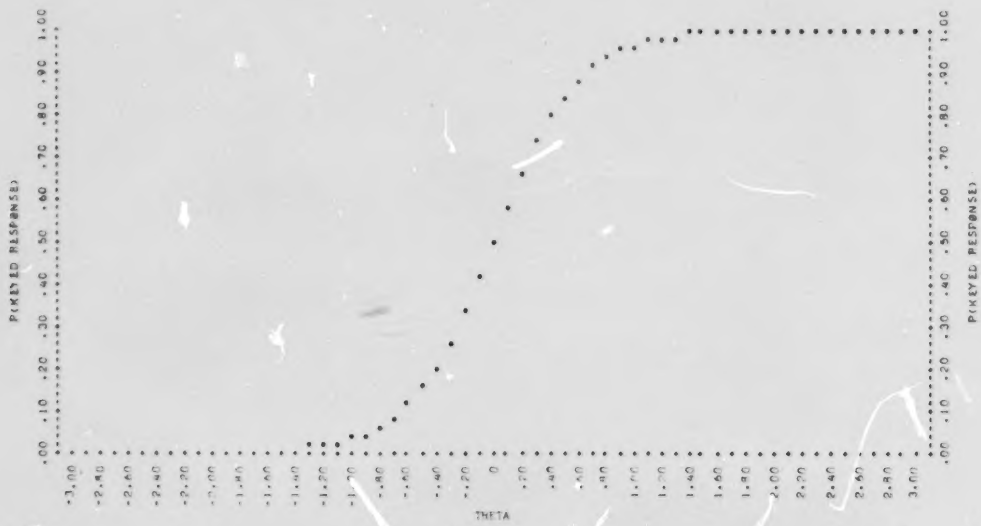
The three-parameter logistic model. The latent trait model that I will use in my example has been investigated extensively by Birnbaum (1968). This model is referred to as the *three-parameter logistic* model, and the function rule that relates probability of a keyed response to the parameters of the model is given in Equation 1.

$$P_g(\theta) = c_g + (1-c_g)[1 + \exp(-1.7a_g(\theta - b_g))]^{-1} \qquad [1]$$

The quantity $P_g(\theta)$ is the probability of a keyed response to item g, with parameters a_g, b_g and c_g, by a person whose location on the latent trait continuum is given by the quantity θ (theta). The exponential operator (exp) indicates that the quantity in parentheses is an exponent of the constant e=2.71828.

Figure 1 shows a graph of the function $P_g(\theta)$ in the interval from θ=-3.00 to θ=+3.00 for an item having a_g=2.0, b_g=0.0, and c_g=.00. This graph was generated by evaluating $P_g(\theta)$ at 61 points along the theta continuum. The irregularities visible in Figure 1 result from rounding $P_g(\theta)$ to the nearest .02 for plotting purposes.

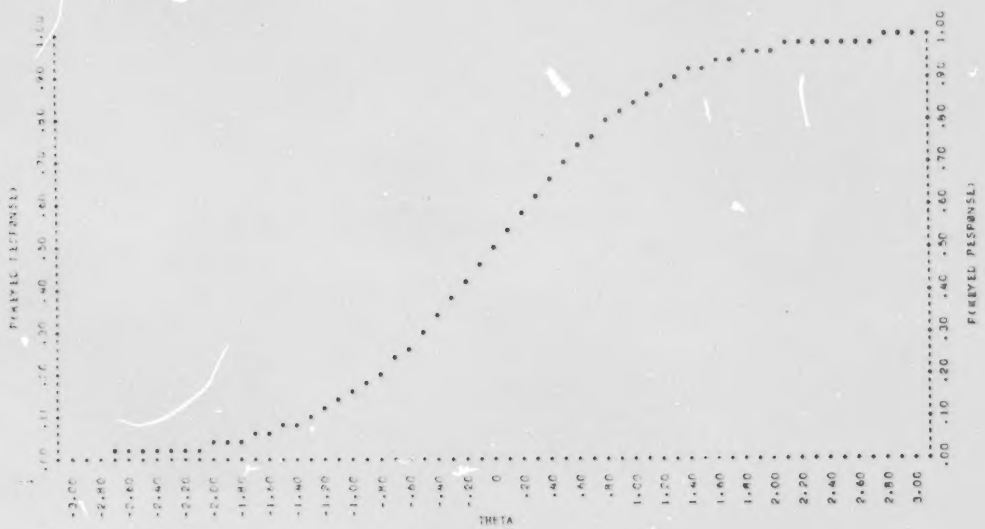
Figure 1
Response Characteristic Curve (a=2.0, b=0.0, c=.00)



The item parameter b_g is known as the item *difficulty* parameter. It is equal to the location on the latent trait continuum at which $P_g(\theta)$ is equal to $.5(1+c_g)$. The item parameter a_g is known as the item *discrimination* parameter. It is related to the slope of the response characteristic curve and in this model is equal to the reciprocal of the distance that one must move along the theta continuum in order to increase $P_g(\theta)$ from $.5(1+c_g)$ to approximately $(.8455(1-c_g))+c_g$. Since a_g=2.0 and c_g=.00 in Figure 1, the distance between the locations on the theta continuum at which $P_g(\theta)$=.5 and $P_g(-)\doteq$.84 is equal to $1/a_g$=.50 theta units. The item parameter c_g is the value of $P_g(\theta)$ when θ=$-\infty$. It is the lower asymptote of $P_g(\theta)$ and is usually conceived of as the probability of a keyed response occurring "by chance" when θ=$-\infty$.

Figure 2 shows a response characteristic curve for an item having a_g=1.0, b_g=0.0, and c_g=.00. The reduced value of a_g, relative to Figure 1, is reflected in the shallower slope of this graph and in

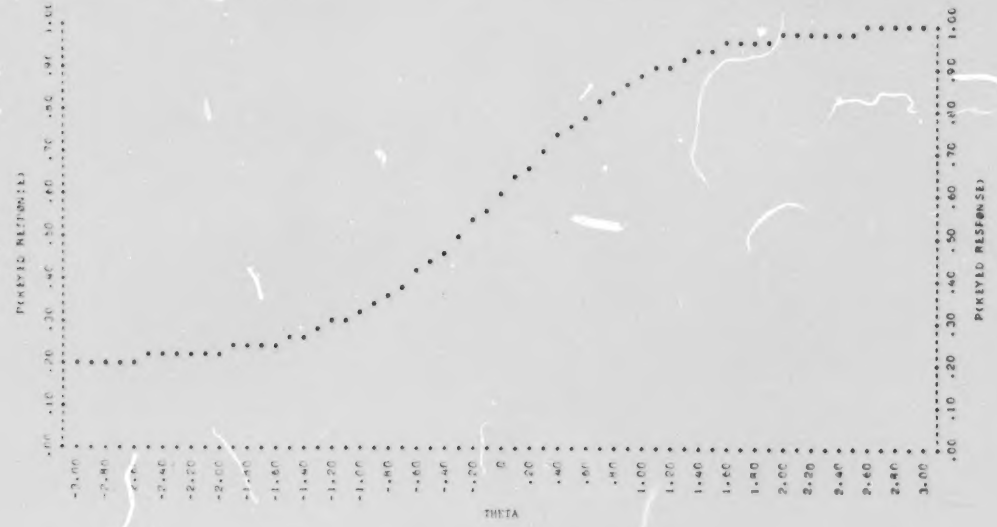
Figure 2
Response Characteristic Curve (a=1.0, b=0.0, c=.00)



the fact that the distance between the locations at which $P_g(\theta)$=.50 and $P_g(\theta)$=.84 is now equal to $1/a_g$=1.00 theta units. A value of a_g in the vicinity of 1.0 is typical of many test items. Values of a_g below about .5 are indicative of "poor" items and values of a_g above 2.0, while desirable in many applications, are not common.

Figure 3 shows a response characteristic curve for an item having a_g=1.0, b_g=0.0, and c_g=.20. The value c_g=.20 might be applicable to a multiple-choice test item that has five response alternatives. In accord with the definitions given above, we see that b_g is equal to the location at which $P_g(\theta)$=.5(1+.2)=.60 and that a_g is equal to the reciprocal of the distance from the location at which $P_g(\theta)$=.60 to the location at which $P_g(\theta)$≐(.8455(1-.2))+.2≐.88. Note that one of the effects of a non-zero c_g is to reduce the slope of $P_g(\theta)$ at all points along the theta continuum.

Figure 3
Response Characteristic Curve (a=1.0, b=0.0, c=.20)



The concept of "information". Birnbaum (1968) has discussed the concept of "information" available in a test item. Birnbaum's *item information function* is given in Equation 2.

$$I(\theta, u_g) = [P'_g(\theta)]^2/[P_g(\theta)\ Q_g(\theta)] \qquad [2]$$

In this equation, u_g is the *item response variable*. It is equal to 1 when a keyed response is emitted and is equal to 0 otherwise. The quantity $Q_g(\theta)$ is equal to $1-P_g(\theta)$. The numerator of Equation 2 is the squared first derivative (i.e., the squared slope) of $P_g(\theta)$ at a fixed value of θ. The denominator is the variance of the item response variable, u_g, at a fixed value of θ. The quantity $I(\theta,u_g)$ is an index of how well we can discriminate people whose latent trait location equals θ from people at nearby latent trait locations on the basis of responses to item g.

In general, a steeper slope for $P_g(\theta)$ implies greater discriminating power. As we noted earlier, high values of a_g and low values of c_g increase the slope of $P_g(\theta)$ and, hence, the information available from an item. The variance of u_g approaches zero at latent

trait levels that are deviant from b_g and reaches its maximum value at the latent trait level where $P_g(\theta)=.5$. Figure 4 shows a graph of the function $I(\theta,u_g)$ in the interval from $\theta=-3.00$ to $\theta=+3.00$ for an item having $a_g=1.0$, $b_g=0.0$, and $c_g=.00$. This graph was generated by evaluating $I(\theta,u_g)$ at 61 points along the theta continuum and rounding the obtained values to the nearest .02.

Figure 4
Information Curve for a Single Item (a=1.0, b=0.0, c=.00)

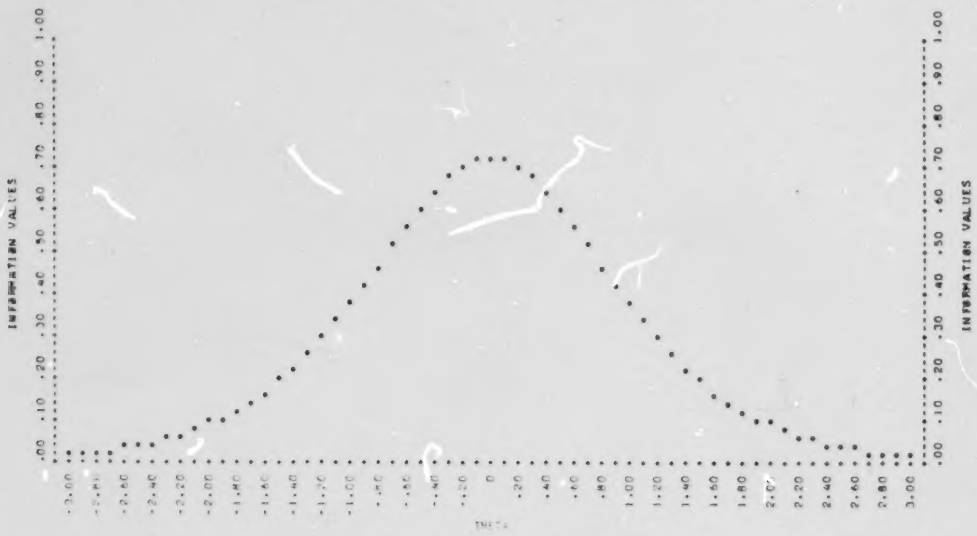


Figure 4 shows that an item provides maximum information in the region of the theta continuum where the item is located (i.e., near b_g) and relatively little information at levels far below or far above b_g. This result is consistent with intuitive impressions of item discriminating power. If, for example, we gave an ability test item that was suitable for third graders (i.e., $P_g(\theta)$ near .5 among third graders) to college students (in which group $P_g(\theta)\doteq 1.0$), we would find that all the college students answered correctly and no basis for discriminating among college students would exist. Note that in Figure 4 the information curve is symmetric about b_g and attains a maximum value of approximately .72.

Figure 5 shows an information curve for an item having $a_g=.85$, $b_g=0.0$, and $c_g=.00$. This curve, while still symmetric about b_g,

attains a lower maximum (approximately .52) and falls off more gradually on either side of b_g than the curve in Figure 4. In fact, the item represented in Figure 5 provides slightly more information than the item represented in Figure 4 in the interval below $\theta \doteq -1.40$ and in the interval above $\theta \doteq 1.40$. However, the gain in these regions is slight compared to the information loss in the interval $-1.40 \leq \theta \leq 1.40$.

Figure 5
Information Curve for a Single Item (a=.85, b=0.0, c=.00)

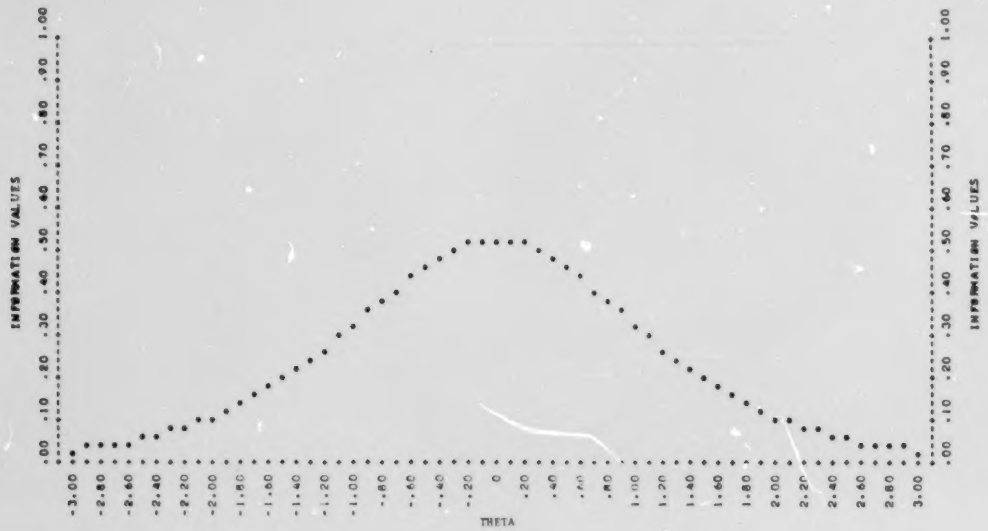
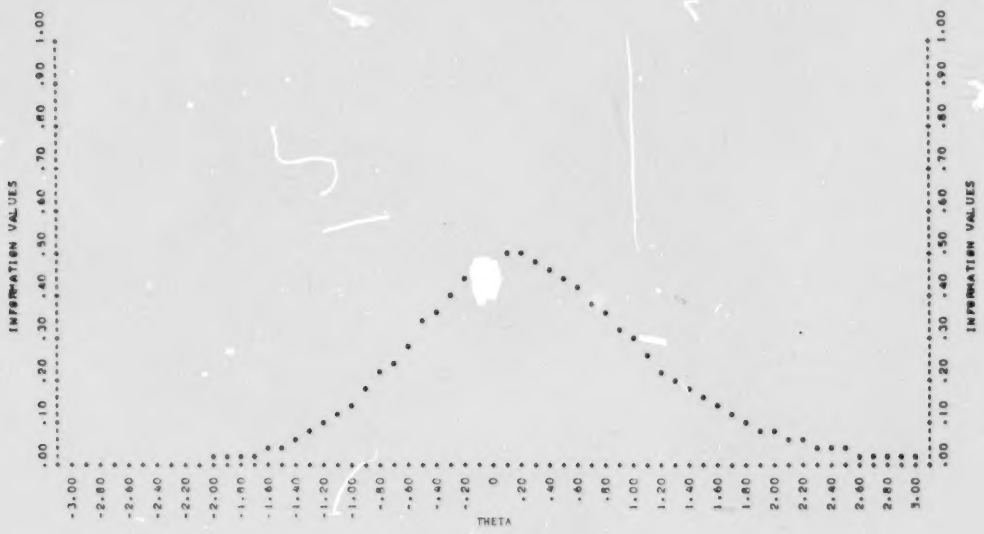


Figure 6 shows an information curve for an item having a_g=1.0, b_g=0.0, and c_g=.20. This curve is not symmetric about b_g. It attains its maximum value of about .50 near θ=.16. The curve falls off more rapidly on the left of θ=.16 than on the right. This reflects the fact that "chance" keyed responses are more prevalent among people located below b_g than among people located above b_g. Such "lucky" responses contribute error to the estimation of latent trait status and reduce the amount of information available. We may note that a_g=1.0, c_g=.20 produces an information curve that is lower than the curve in Figure 5. Thus, introducing the possibility of "lucky" keyed responses reduces the information available from

an item just as if we had selected an item with lower a_g, but with c_g=.00.

Figure 6
Information Curve for a Single Item (a=1.0, b=0.0, c=.20)



Sequential estimation in an adaptive test. In order to demonstrate the sequential estimation of latent trait status in an adaptive test, a computer program was used to simulate the test responses of a person whose latent trait location is θ=+1.0. A total of 20 items having a_g=1.0 and c_g=.20 was administered. The item b_g values changed from item to item as a function of responses generated during the simulated test. Table 1 summarizes the results of this 20-item test.

The first column in Table 1 contains item numbers in the 20-item series (g=1,2,...,20). The second column contains the b_g values of the items administered. The difficulty of the first item was b_1=0 because this value approximates the mean latent trait score in any population of persons that is sampled to parameterize a set of test items. (An exception to this may be found in Wright & Panchapakesan's (1969) implementation of the Rasch model. They scale the latent trait metric such that the mean of the b_g estimates is zero and the mean θ estimate among persons is, in general, other than zero.) Following the first item, b_g values either increase or decrease (in

accordance with a procedure that I will outline in a moment) depending on whether a keyed or non-keyed response was generated. The item response variable u_g is shown in the third column of Table 1.

Table 1
Sequential Estimation of Latent Trait Status in a 20-Item Adaptive Test

Item No.	Diff.	Resp.	MAXL. Est.	W.B.L. Est.	SBAYES Est.	OBAYES Est.
1	0	1	5.49	1.61	.38	.38
2	1.00	0	.36	-.85	.05	.04
3	0	1	.67	.18	.32	.31
4	.18	1	.89	.82	.53	.54
5	.82	1	1.16	1.25	.75	.78
6	1.25	0	.87	.72	.57	.56
7	.72	1	1.03	1.00	.74	.75
8	1.00	1	1.20	1.21	.89	.93
9	1.21	0	.99	.93	.74	.74
10	.93	1	1.12	1.10	.87	.89
11	1.10	0	.95	.89	.73	.72
12	.89	1	1.05	1.02	.84	.84
13	1.02	0	.91	.85	.72	.70
14	.85	1	.99	.96	.82	.80
15	.96	1	1.07	1.05	.90	.90
16	1.05	0	.96	.92	.80	.78
17	.92	1	1.03	1.00	.88	.87
18	1.00	0	.93	.89	.79	.76
19	.89	1	.99	.96	.86	.84
20	.96	1	1.05	1.03	.92	.92

The last four columns of Table 1 contain four different estimates of latent trait status that were calculated after each item was administered. The fourth column contains (approximate) maximum-likelihood estimates of θ. A maximum-likelihood estimate of θ corresponds to the latent trait location at which the observed item responses have maximum probability of occurrence. The probability of a set of item responses, given some fixed value of θ and the item parameters, is obtained using the *likelihood function* given in Equation 3.

$$L_k(\theta) = \prod_g [P_g(\theta)^{u_g} Q_g(\theta)^{1-u_g}] \qquad [3]$$

This equation assumes that the responses of a given person to different test items are independent of one another. The operator $\prod$ indicates that a serial product is to be taken over the test items administered up to that point ($g=1,2,\ldots,k$).

After each item was administered, Equation 3 was evaluated at 101 equally spaced θ values in the interval from θ=-5.00 to θ=+5.00 and the largest of the 101 likelihood values was identified. Then, a quadratic function was fitted to this largest likelihood value and the two likelihoods adjacent to it. The value of θ corresponding to the maximum of the quadratic function was used as the "MAXL." estimate. Under most conditions, the estimate of θ obtained in this manner is a good approximation to the estimate that would be obtained if we used methods of numerical analysis to search for a root of the log-likelihood function's first derivative. The procedure described here has the advantage that it is not subject to local maxima in the interval between θ=-5.00 and θ=+5.00.

As an aside, we may note that the interval between -5.00 and +5.00 will contain at least 96% of the θ estimates in any group that is used to parameterize test items. This is because latent trait item parameterization procedures scale the theta metric such that the mean θ estimate equals zero and the standard deviation among the estimates is 1.0 (again, the Rasch model provides an exception to this general result), and by virtue of Tchebycheff's inequality which states that the proportion of cases which fall more than S standard deviations from the mean cannot exceed $(1/S^2)$ in any distribution (Hays, 1973, p. 253). If the distribution is peaked and unimodal, virtually all of the θ estimates will be between -5.00 and +5.00.

Figure 7
Relative Likelihood and Posterior Probability Curves After 1 Item

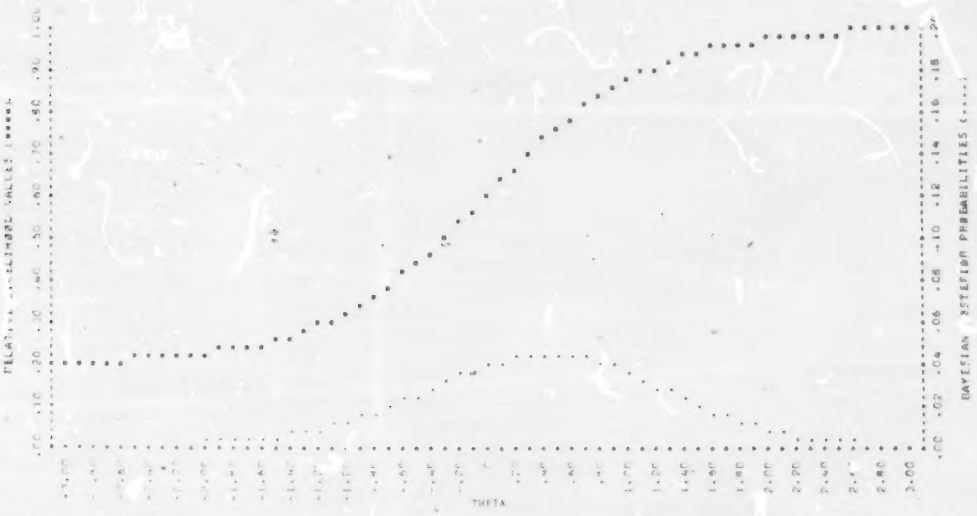


Figure 8
Relative Likelihood and Posterior Probability Curves After 2 Items

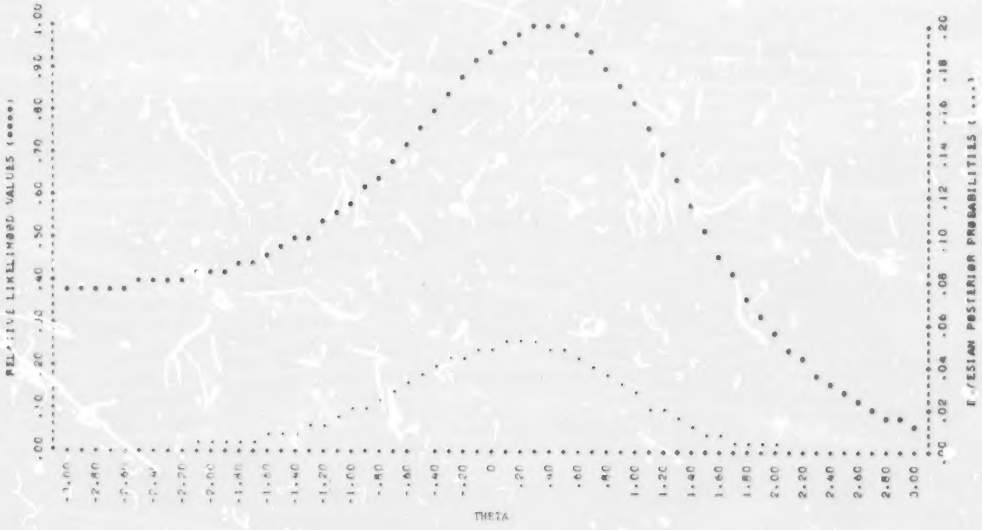
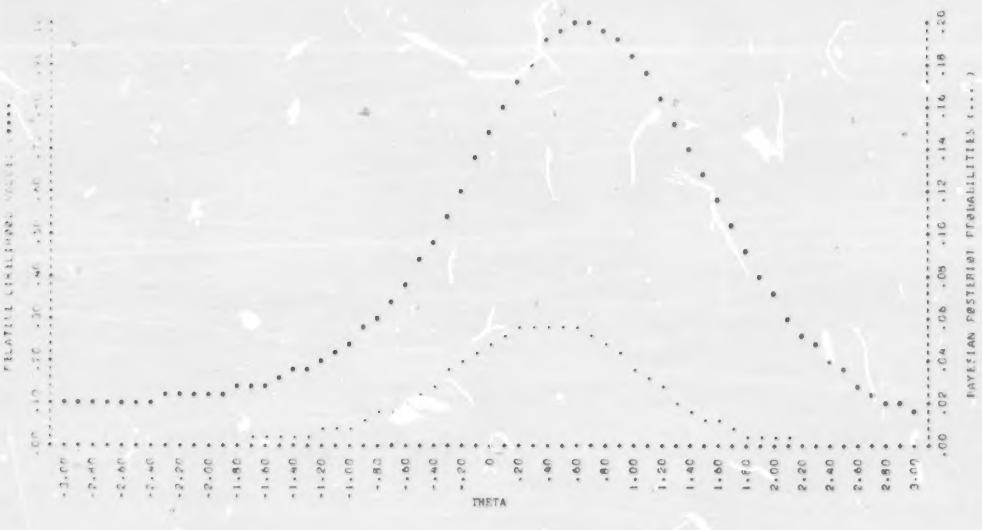


Figure 9
Relative Likelihood and Posterior Probability Curves After 3 Items



Figures 7, 8, and 9 show graphs of the data likelihood function in the interval from θ=-3.00 to θ=+3.00 following the administration of 1, 2, and 3 items, respectively. For plotting purposes, the raw likelihood values were expressed relative to the largest likelihood value in the interval θ=-5.00 to θ=+5.00 and then rounded to the nearest .02. As can be seen in Equation 3, after one item is administered the likelihood function corresponds to either $P_j(\theta)$ or $Q_j(\theta)$, depending on whether a keyed or non-keyed response is emitted. (Compare Figure 7 and Figure 3.) The MAXL. estimate after getting the first item "correct" is +5.49. Actually, since $P_g(\theta)$ is strictly increasing in θ, the estimate should be θ=+∞, but a finite estimate is certainly more reasonable. After "missing" the second item, with b_g=1.00, the peak of the likelihood curve occurs near θ=+.36 (Figure 8). After the third item, the peak occurs near θ=.67 (Figure 9).

A "weighted-by-likelihoods" (W.B.L.) estimate of latent trait status appears in the fifth column of Table 1. The W.B.L. estimates were obtained by taking a weighted average of 101 equally spaced θ values in the interval from θ=-5.00 to θ=+5.00. The weights used were the data likelihoods at each θ value. That is,

$$\text{W.B.L. EST.} = [\sum_{\theta}(L_v(\theta)\ \theta)]/[\sum_{\theta}(L_v(\theta))] \qquad [4]$$

where θ takes on the values -5.00, -4.90, ..., +5.00. The W.B.L. estimate is influenced by the entire set of 101 likelihood values instead of just the three largest values.

The MAXL. and W.B.L. estimates can differ considerably when only a few items have been administered, as can be seen in Table 1. Inspection of the relative likelihood curve in Figure 8 shows why these two estimators differ after two items have been administered. The W.B.L. estimate is lower due to the fact that the left tail of the likelihood curve is high relative to the right tail. Table 1 also shows that the MAXL. and W.B.L. estimators become more similar as the number of items administered increases. To my knowledge, the W.B.L. estimator has not been proposed previously. I plan to study its characteristics in future research.

At this point I can outline the procedure by which item b_g values were determined during the simulated test. The general rule followed was: Let the next item have a difficulty level equal to the current value of the W.B.L. estimator, except that in no case shall the new b_g value be more than 1.00 units from the immediately preceding b_g value. Thus, as can be seen in Table 1, item difficulties changed by 1.00 until the third item had been administered and the W.B.L. estimate was .18. After this, each item difficulty corresponded to the value of the W.B.L. estimate following the preceding item. In actual practice we seldom find an item with b_g exactly equal to the

current estimate of latent trait status. In such cases we select an item that has b_g close to the desired value.

Columns six and seven of Table 1 contain two Bayesian estimates of latent trait status. Given a specified form for the continuous distribution of latent trait scores in a population (i.e., the *prior probability density function* of theta) the item parameters for the items administered, and a vector of item responses (u_g values), one can, in principle, derive the *posterior probability density function* of theta using the inverse probability rule of Bayes (Hays, 1973, p. 819). In practice, it becomes difficult to obtain analytic expressions for the posterior theta distribution unless the prior distribution and the data likelihood function take on certain restricted forms. To avoid such difficulties, I have proposed the following approximate procedure.

First, we approximate the continuous prior density function of theta with a discrete probability distribution in which the probabilities are concentrated at 101 equally spaced points along the theta continuum. Thus, for example, the area under the prior density curve between θ=-.05 and θ=+.05 is assigned to the point θ=.00. This is done for θ=-5.00, -4.90, ..., +5.00. Areas beyond θ=-5.05 and θ=+5.05 are assigned to the points θ=-5.00 and θ=+5.00 respectively. (These extreme tail areas should be trivially small. If they are not, one can either shift or extend the region of the theta continuum in which the procedure is applied.) Next, data likelihoods are generated at the same 101 values of θ using Equation 3. The prior probabilities, $f(\theta)$, and the data likelihoods, $L_v(\theta)$, are then entered into Equation 5 in order to determine the posterior probability of each given θ value.

$$P(\theta|v) = [L_v(\theta)\ f(\theta)]/\sum_{\theta}[L_v(\theta)\ f(\theta)] \qquad [5]$$

The resulting 101 posterior probabilities provide a discrete approximation to the continuous posterior distribution of theta. Finally, the mean of the discrete posterior distribution is obtained with Equation 6 and this value is referred to as the "SBAYES" (simplified Bayesian) estimate at that stage of the testing procedure.

$$\text{SBAYES } \hat{\theta} = \sum_{\theta}[P(\theta|v)\ \theta] \qquad [6]$$

Figures 7, 8, and 9 show three of the posterior probability distributions that were generated with the SBAYES procedure when the prior distribution of latent trait scores was specified to be a normal density function with zero mean and unit variance. The first three SBAYES estimates in Table 1 are the means of these discrete distributions.

The "OBAYES" (Owen Bayesian) latent trait estimates that appear in column seven of Table 1 were obtained using a procedure described

by Owen (1975). While Owen described a method for estimating latent trait status and a method for selecting test items, I have utilized only his estimation procedure here. Owen introduced his procedure in the context of a *three-parameter normal ogive* latent trait model. The close similarity of this model to the logistic model given in Equation 1 allows its application here.

The OBAYES procedure has two drawbacks. First, it is limited to prior distributions that follow a normal density function. The SBAYES procedure described above can accept any type of prior distribution. Second, the OBAYES procedure is order dependent. That is, if a set of items is administered and the item responses are recorded, then the value of the OBAYES estimator will depend partly on the order in which the items are processed by the scoring procedure. The OBAYES procedure (implicitly) generates an updated prior distribution after each item is scored and then combines this new prior distribution with the likelihood function for the response to the next item. This in itself would not make the OBAYES procedure order dependent but, in order to simplify the mathematics, Owen proceeded as if each updated prior distribution could be described by a normal density function. This approximation introduces a small amount of inaccuracy into the estimation process and makes the procedure order dependent. The SBAYES procedure utilizes the same prior distribution at each stage of the testing process and updates the data likelihood function after each item is scored. The SBAYES procedure is not order dependent.

After administering a single item, SBAYES and OBAYES estimates generally agree to three decimal places when the prior distribution is a normal density function. Since the OBAYES estimate is optimal in this particular situation, we can view this level of agreement as an indication that very little inaccuracy is introduced by the discrete approximations in the SBAYES procedure. When more than one item has been administered, or when the prior distribution specified for the SBAYES procedure is non-normal, the two estimation methods will not necessarily agree, as can be seen in Table 1.

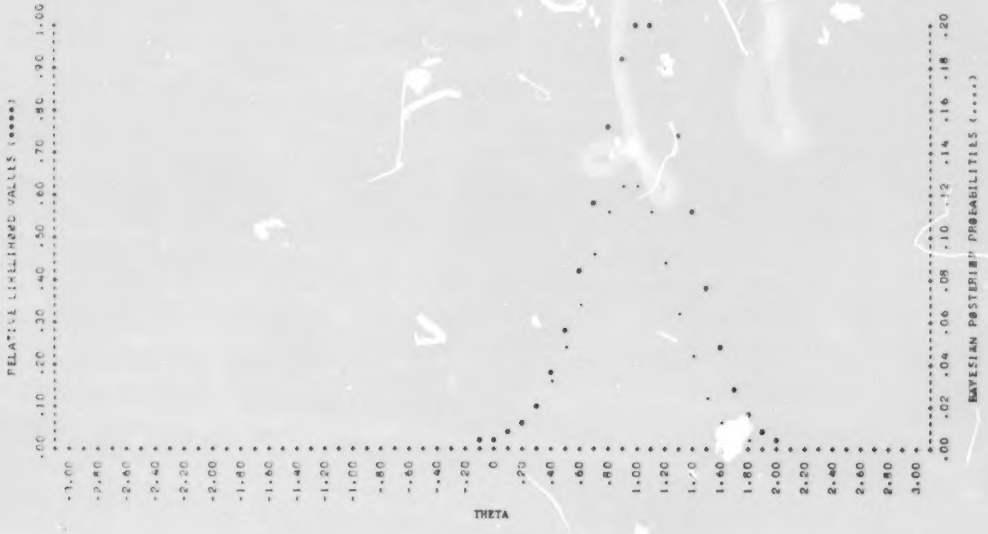
Figure 10 shows the relative likelihood and posterior probability curves that resulted after 20 items had been administered. The likelihood curve peaks near $\hat{\theta}$=1.05 and the posterior probability distribution has a mean of .92 (see Table 1). Both the likelihood curve and the posterior probability curve have shifted to the region of the theta continuum near θ=1.00, and both curves have become more peaked. In fact, as test length (k) approaches infinity, both of these curves approach a vertical line (i.e., a single-valued distribution) located at the value of θ that is generating the item responses.

Note in Table 1 that the Bayesian estimates of θ tend to stay closer to θ=.00 than the other estimates throughout the testing process. This is because Bayesian estimators are "drawn toward" the high density region of the prior distribution. This is appropriate when one's objective is to minimize squared errors of estimation in the population specified by the prior distribution. Unfortunately, for tests of moderate length, we must accept a certain amount of

bias at the tails of the theta distribution in order to achieve this minimization (McBride & Weiss, 1976).

For moderate k, the maximum-likelihood estimator can also be biased. However, for a given value of k and values of θ deviant from the high density region of a peaked prior distribution, the maximum-likelihood estimator will tend to be less biased than the Bayesian estimator. The Bayesian estimator's bias can be reduced by increasing k as the estimate of θ deviates from the high density region of the prior distribution. This can be done readily in an adaptive testing situation.

Figure 10
Relative Likelihood and Posterior Probability Curves After 20 Items



An interesting relationship exists between the MAXL. and W.B.L. estimators and Bayesian estimators. If one applied the SBAYES estimation procedure and specified that the prior distribution of theta was rectangular in the interval θ=-5.05 to θ=+5.05, then the SBAYES estimate of θ, as determined by Equation 6, would be identical to the W.B.L. estimator. Moreover, the MAXL. estimate would closely approximate the mode of the Bayesian posterior probability distribution. Thus, we see that all four types of latent trait estimator that have been presented here can be viewed as Bayesian estimators. The MAXL. estimator is a Bayesian modal estimate of θ when the (implicit) prior is restricted to a rectangular form, the W.B.L. estimator is a least-squares estimate of θ when the (implicit) prior is restricted to a rectangular form, and the OBAYES estimator is a least-squares estimate of θ when the (explicit) prior is restricted to a normal form. The SBAYES procedure is the only one of the four

methods that does not restrict the form of the prior distribution. By virtue of this flexibility, I am inclined to recommend the SBAYES estimation procedure as the most widely applicable of the four methods.

Total test information. Birnbaum (1968, p. 454) has defined the *information function of a test* as

$$I(\theta) = \sum_g [I(\theta, u_g)]. \qquad [7]$$

This function is the sum of the constituent item information functions and defines the maximum amount of information that can be extracted from a set of items. The amount of information actually extracted depends on how the items are scored.

Figure 11 shows a graph of the test information function for the 20 items administered in our simulated adaptive test. It was obtained by evaluating Equation 7 at equally spaced points along the theta continuum in the interval from θ=-3.00 to θ=+3.00. This curve shows the maximum amount of information available from the set of items, regardless of how they are scored. The curve peaks near θ=1.00, thus indicating that this set of items provides maximum discrimination between individuals whose latent trait locations fall near θ=1.00. The maximum value of the curve is about 9.00.

Figure 11
Information Curve for 20-Item Adaptive Test

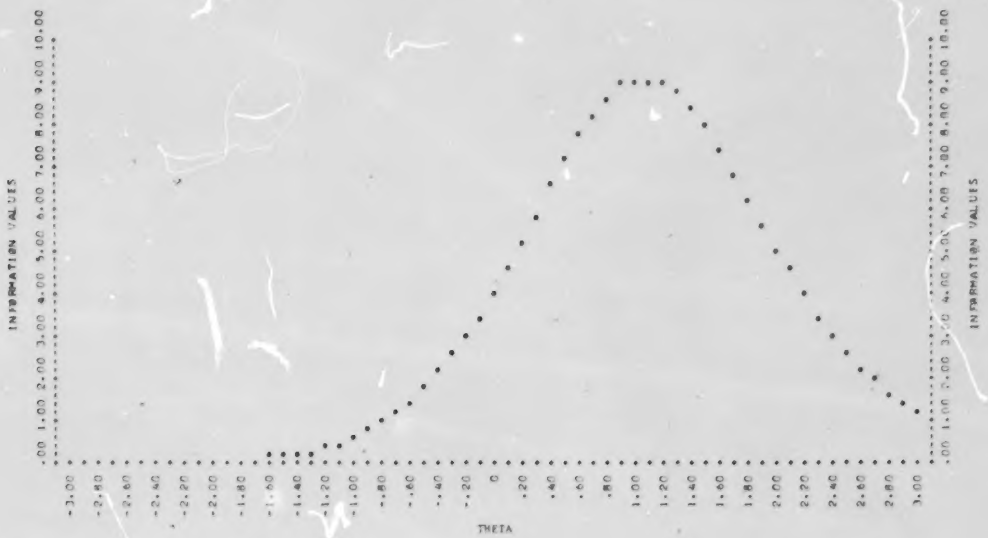


Figure 12
Information Curve for 20-Item Rectangular Test ($-3.0 \leq b \leq +3.0$)

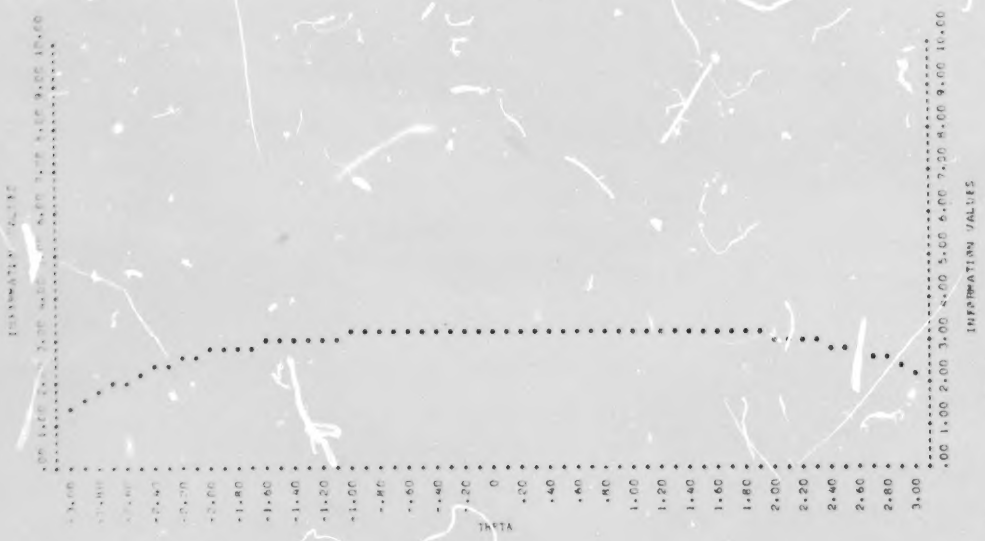


Figure 13
Information Curve for 20-Item Peaked Test (b=0.0 for all items)

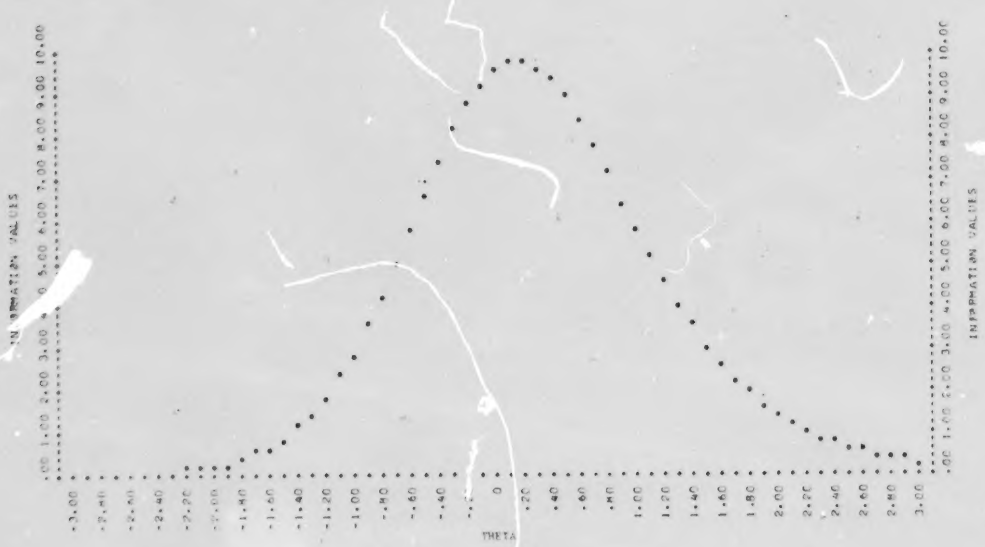


Figure 12 shows a graph of the test information function for a set of 20 items having a_g=1.0, c_g=.20, and b_g values equally spaced in the interval from -3.00 to +3.00 (i.e., b_g=-3.00, -2.68, -2.37, ..., +3.00). This would commonly be referred to as a "rectangular" distribution of item difficulties. This test provides a fairly uniform level of information across a broad range of the theta continuum. Unfortunately, the level of information is relatively low. The curve attains its maximum value of about 3.20 in the interval $-1.00 \leq \theta \leq 1.90$.

Figure 13 shows a graph of the test information function for a set of 20 items having a_g=1.0, c_g=.20, and b_g=0.0 for all items. This is a "perfectly peaked" test. The shape of this information curve is rather similar to the curve in Figure 11, but it is shifted to the left. The curve in Figure 13 attains its maximum value of 9.80 near θ=.16. At θ=1.00, the value of this information curve is about 5.80.

Figures 12 and 13 represent two rather idealized non-adaptive tests. We have seen that both of these tests deliver less information at θ=1.00 than the items selected by the adaptive testing procedure. What is the implication of this result? If we had to order our imaginary testee relative to other individuals falling at nearby θ values, we would make fewer mistakes if we used θ estimates derived from the adaptive test's items than if we used estimates derived from either conventional test's items.

Concluding Remarks

I have presented several procedures for estimating latent trait status. I have also suggested that adaptive testing procedures often can provide more accurate estimates of latent trait status than conventional tests. Though there is no necessary connection between latent trait theory and adaptive testing, there is a strong natural impetus toward their joint application. Latent trait theory provides adaptive testing with a coherent theoretical foundation. It is a guide to procedures for designing and scoring adaptive tests. On the other hand, adaptive testing offers us the opportunity to take maximum advantage of the potentialities of latent trait theory. Adaptive testing offers the possibility that test reliability and/or validity might be increased or that some specified level of reliability or validity might be achieved in less testing time. We are now at a fortunate point in time where both a new type of test theory and a new type of testing technology are available. Their joint effect might possibly exceed the sum of the two parts.

REFERENCES

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968, Chapters 17-20.

Bock, R.D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 1972, 37(1), 29-51.

Christoffersson, A. Factor analysis of dichotomized variables. Psychometrika, 1975, 40(1), 5-32.

Cronbach, L.J. Test validation. In R.L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Glaser, R., & Nitko, A.J. Measurement in learning and instruction. In R.L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Hambleton, R.K., & Novick, M.R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10(3), 159-170.

Hays, W.L. Statistics for the social sciences. New York: Holt, Rinehart, & Winston, 1973.

Lord, F.M. A theory of test scores. Psychometric Monograph, 1952, No. 7.

Lumsden, J. Test theory. Annual Review of Psychology, 1976, 27, 251-280.

McBride, J.R., & Weiss, D.J. Some properties of a Bayesian adaptive ability testing strategy. (Research Report 76-1). Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, 1976.

Nunnally, J.C. Psychometric theory. New York: McGraw-Hill, 1967.

Owen, R.J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70(350), 351-356.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Neilson & Lydiche (for Danmarks Paedagogiske Institut), 1960.

Samejima, F. Normal ogive model on the continuous response level in the multidimensional latent space. Psychometrika, 1974, 39(1), 111-121.

Wright, B.D., & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.